

**A COMPARATIVE STUDY ON DETECTING BANGLA FAKE NEWS ON SOCIAL
MEDIA USING MACHINE LEARNING ALGORITHMS**

BY

**ADIL ANSARY
ID: 172-15-10104**

AND

**MOHAMMAD RAKIB HASSAN
ID: 172-15-9794**

This Report Presented in Partial Fulfillment of the Requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Supervised By

Zerin Nasrin Tumpa

Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

MD. AZIZUL HAKIM

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

SEPTEMBER 2021

APPROVAL

This Project titled “**A comparative study on detecting Bangla fake news on social media using machine learning algorithms**”, submitted by Adil Ansary, ID No: 172-15-10104 and Mohammad Rakib Hassan, ID No: 172-15-9794 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 09/09/2021.

BOARD OF EXAMINERS

Chairman



Dr. Touhid Bhuiyan
Professor and Head

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

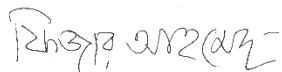
Internal Examiner



Nazmun Nessa Moon
Assistant Professor

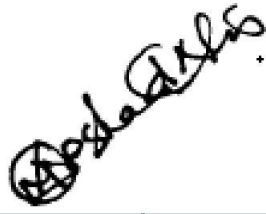
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner



Dr. Fizar Ahmed
Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



External Examiner

Dr. Md Arshad Ali
Associate Professor
Department of Computer Science and Engineering
Hajee Mohammad Danesh Science and Technology
University

DECLARATION

We hereby declare that this thesis has been done by us under the supervision of **Ms. Zerine Nasrin Tumpa**, Lecturer, **Department of CSE**, and co-supervision of **Mr. Md. Azizul Hakim**, Lecturer, **Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of any degree or diploma.

Supervised by:



Ms. Zerine Nasrin Tumpa

Lecturer

Department of CSE

Daffodil International University

Co-Supervised by:



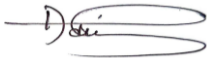
Mr. Md. Azizul Hakim

Lecturer

Department of CSE

Daffodil International University

Submitted by:



Adil Ansary

ID: 172-15-10104

Department of CSE

Daffodil International University



Mohammad Rakib Hassan

ID: 172-15-9794

Department of CSE

Daffodil International University

ACKNOWLEDGEMENT

First of all, we want to render our gratitude to the Almighty Allah for the enormous blessing that makes us able to complete the final thesis successfully.

We are really grateful and express our earnest indebtedness to **Ms. Zerin Nasrin Tumpa**, Lecturer, Department of CSE Daffodil International University, Dhaka, Bangladesh. Profound Knowledge & intense interest of our supervisor in the field of “Machine Learning” make our way very smooth to carry out this thesis. Her remarkable patience and dedication, scholarly guidance, continual encouragement, vigorous motivation, direct and fair supervision, constructive criticism, valuable advice, great endurance during reading many inferior drafts and correcting the work to make it unique paves the way of work very smooth and ended with a great result. We would like to express our gratitude wholeheartedly to **Prof. Dr. Touhid Bhuiyan**, Professor, and Head, Department of CSE, for his kind help to finish our thesis and also to other faculty members and the staff of CSE department of Daffodil International University. We would like to express thankfulness to the fellow student of Daffodil International University, who took part in this discussion during the completion of this work. Finally, we must acknowledge with due respect the constant support and passion of our parents and family members.

ABSTRACT

One of the most innovative inventions of our time is social media. It is extremely important for each of us, with its own mix of benefits and drawbacks.

Fake News has now become a big issue, causing chaos all over the globe. As a result, developing an algorithm that is as accurate as possible would be a revelation, with far-reaching implications for social problems that are widespread, also the ongoing political situation. People use social networks also online based news articles as a primary basis of the information and news because they are easy to access, have a low cost, and are easy accessible, just a click away. But it has a number of drawbacks, including no way to verify the source, reliability, or validity of the viewpoints being endorsed.

The veracity of the perspectives being supported. As a result, we've used four machine learning algorithms (Logistic regression, Decision tree classification, Gradient Boosting Classifier, Random Forest Classifier) for comparison to see which algorithm is more effective and efficient and better for predicting the result.

TABLE OF CONTENTS

CONTENTS	PAGE
Board of examiners	i
Declaration	iii
Acknowledgements	iv
Abstract	v
List of Figure	viii
List of Tables	ix
 CHAPTER	
 CHAPTER 1: INTRODUCTION	1-3
1.1 Introduction	1
1.2 Motivation	2
1.3 Research Questions	2
1.4 Expected Output	2
1.5 Report Layout	3
 CHAPTER 2: BACKGROUND STUDIES	5-6
2.1 Introduction	5
2.2 Related Works	5-6
2.3 Research Summary	6
2.4 Scope of the Problem	6
2.5 Challenges	6
 CHAPTER 3: RESEARCH METHODOLOGY	7-15
3.1 Introduction	7

3.2	Research subject and instrumentation	7
3.3	Data Collection procedure	7-8
3.4	Statistical analysis	8
3.5	Algorithms and procedure	8-15
 CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION		 16-19
4.1	Introduction	16
4.2	Experimental results	16-19
4.3	Summary	19
 CHAPTER 5: CONCLUSION AND FUTURE WORK		 20-21
5.1	Summary of the study	20
5.2	Conclusion	20
5.3	Recommendations	21
5.4	Further Study	21
 REFERENCES		 22
 APPENDICES		 25-35

LIST OF FIGURES

FIGURES	PAGE
Figure 1.4.1 Sample Output	3
Figure 3.4.1 Statistics of dataset	8
Figure 3.5.1 RFC diagram	13
Figure 3.6.1 Framework	14
Figure 3.6.2 Used Libraries	15
Figure 4.2.2 Pie chart	17
Figure 4.2.3 Confusion Matrix	17
Figure 4.2.4 Confusion Matrix	18
Figure 4.2.5 Confusion Matrix	18
Figure 4.2.6 Confusion Matrix	19

LIST OF TABLES

TABLE	PAGE
Table 3.5.1 Logistics Regression Classification Report	11
Table 3.5.2 Decision Tree Classification Report	11
Table 3.5.3 Gradient Boost Classification Report	12
Table 3.5.4 Random Forest Classification Report	13
Table 4.2.1 Accuracy Table	16

CHAPTER 1

INTRODUCTION

1.1 Introduction

Fake news could be described as some false news fabricated as true news. Sarcastic news, hoaxes, totally fabricated news, and government propaganda are the most common examples of fake news. It's usually spread to draw in viewers and raise advertising revenue. The reason of increasing fake news are media distortion and misinformation, governmental and social power, incitement and social conflict, and economical benefit [1]. People and organizations using possibly malicious motives, on the other hand, likely be known to spread false news with the thought to manipulate occasions and policies all over the world. [2] shows us how fake news had a huge effect on the 2016 United States presidential elections. The spread of myths and false information have reached a breaking point in recent years, to the point that it is now affecting social and politically aware issues also. The times that people spend on social networks and online based news sources is rapidly increasing. As the result, they get the majority of their information from social media sources. While social media can be used from everywhere all the time and is open, it offers anonymity when sharing one's opinion, resulting in a lack of accountability. As opposed to a paper or some other reliable news sources, this significantly decreases the reliability of data obtained from them. Due to a lack of continuous monitoring and oversight, miscreants have been able to run wild and spread false information. [3] Analyzed the behaviors of respondents' social media profiles to perform a study about how fake and manipulated news is distributed on online media platforms alike "Facebook". The World Economic Forum reported in 2013 that the so-called "digital wildfires", or untrustworthy news spreading virally online (also known as fake news), will be one of society's greatest challenges. Given its negative consequences (for example, it causes unnecessary chaos and social unrest in society) [4][5], the identification of fake news is becoming more and more essential [6]. Researchers have recently been very interested in identifying bogus news. When question is to identify fake manipulated news, various methods have been used so far [7].

According to a new survey, nearly one-third of people in Spain, the United States, South Korea, Germany, Argentina, and the United Kingdom claim to have seen misleading or inaccurate information about COVID-19 on social media [8]. Many critics are now denoting the current upsurge to fabricated news related to the COVID-19 widespread as disinformation due to the widespread dissemination of inaccurate and unreliable details [9]. Misusing the news that is broadcast to different consumers will only lead to uncertainty and confusion because there will be several versions of the facts.

Our goal is to create a model that will predict a news is true or fake using machine learning algorithms with an increased chance to stop negativity and confusion across the country. We are going to use four Machine learning algorithms (Logistic regression, Decision tree classification, Gradient Boosting Classifier, Random Forest Classifier) and will compare results based on the accuracy. Beside there has been a little of work done Bangla news so we think our research will be a huge contribution to our country.

1.2 Motivation

Fake news is like cancer, it spreads at a very speed rate. As we know hundreds of conflict happened in our country including the Ramu, Cox's bazar Buddhist temple incident was also happened because of a fake news. Later he was arrested for making the fake news [10] The motivation behind is to ensure decrease fake news as much as possible to avoid casualties happening around our country for fake news.

1.3 Research Questions

While doing the research these are some questions that hit my mind. These are:

1. What is fake news?
2. How fake news spread so fast?
3. Why someone starts spreading fake news?
4. How they make it look like real?
5. What's the purpose of them?

6. How they make it interesting?

1.4 Expected Output

In this research, we tried to develop a model for predicting a news is whether fake or real using some different Algorithms of Machine Learning. Later we will compare them to see which one is more efficient and which one is not.

The outcome will look like this:

```
প্রেম করলে বাড়বে ওজন!  
  
LR Prediction: Fake News  
DT Prediction: Fake News  
GBC Prediction: Fake News  
RFC Prediction: Fake News
```

Fig 1.4.1: Output

1.5 Report Layout

Chapter 1: Introduction

In this part, we've discussed about the motivation to do this research and also the research questions and the anticipated outcome of the research.

Chapter 2: Background

In this chapter, we have discussed about the related work of our research and comparative studies between our developed model and existing related model as well as the problems and challenges that we've faced.

Chapter 3: Research Methodology

We have discussed about our research methodology that means we will talk about how we did our research and about the terminology.

Chapter 4: Experimental results and Discussion

We have discussed about our research experimental results and outputs. Also about which results we have found by implementing the algorithms of machine learning over our collected data in this part.

Chapter 5: Summary, Conclusion, Recommendation and Implication for Future Research

In this chapter, we have discussed about summary, conclusion, recommendation and scope for further development of our research

CHAPTER 2

BACKGROUND STUDIES

2.1 Introduction

Machine Learning enables IT frameworks to recognize designs based on previous calculations and data sets and to generate adequate arrangement ideas. Its algorithms are extremely effective at predicting any outcome provided any data or information. As a result, we used machine learning algorithms to make predictions.

2.2 Related Works

To automate the procedure to identify fake news, we need algorithms that can look at a variety of characteristics that fake news has. Using automatic classification methods, the most difficult issue of misleading language detection [11] is discussed. Syntactic stylometry was used in another study [12] to deal with misleading feedback. To enhance review spam detection efficiency, “Hai et al. [13] devised a semi-supervised learning method based on Laplacian regularized logistic regression”. In terms of truth exploration, most previous study has done so by evaluating the reliability of information sources. Those who do so by employing link-based metrics [14, 15,16], Measures based on information retrieval [22], accuracy-based tests [17, 18, 19, 20, 21], Content-based approaches [23, 24, 25, 26, 27, 28, 29], graphical models [30, 31, 32, 33], and survival analysis [34, 35] are all examples of content-based measures. The authors of [36] use syntactic and semantic structures of media articles to distinguish real and fabricated news that used a trigram language model, with a 91.5 percent accuracy. The creation of a fabricated news detector based on social feature of news has also been studied. The writers of [37] demonstrated that Social media posts can be categorized with a high degree of accuracy as they classified them for deceptions or for non-hoaxes established on whether consumers "liked" or "disliked" the news, and they obtained accuracies of more than 99 percent even with a limited training dataset. The writers used crowd-sourcing methods such as “logistic regression and harmonic boolean mark crowd-sourcing”. While the approach suggested by [38] has a high level of accuracy, it is only applicable to cases that receive sufficiently social articles attention (the number of dislikes or likes). The authors of [39] suggested a

“Multi-source Multi-class Fake News Detection (MMFD)” system and an automatic and observable technique to combine data from several sources, but the models accurateness on real world data is only 38.81 percent.

2.3 Research Summary

It is being attempted to create a model for predicting false or real news, which will be used in future years using various and powerful machine learning algorithms. There have been some related works that have used various types of algorithms on different kinds of gathered data and information. Others have used news articles from sources other than social media (Facebook, Twitter) to detect fake news using various methods. In this study, we used news articles and titles from Facebook and Twitter as part of our data set to calculate whether a news is false or trusted using machine learning algorithms like Random Forest Classifier, Logistic regression, Gradient Boosting Classifier, Decision tree classification Following that, we'll manually test a news headline and compare the accuracy rates of various Machine Learning Algorithms.

2.4 Scope of the Problem

There is always the possibility of a problem in any study. As a result, there really is no exception in our study. Our gathered data should be relevant, and our data set should be easy and smooth to work with. When working with algorithms, we must properly train our collected data. If we don't do this correctly, our performance would be inappropriate and unrepresentative of a real-life situation. We must operate correctly and carefully if we are to stop or decrease the problem.

2.5 Challenges

There is no way to complete a task without encountering difficulties. Challenges can arise at any point during the process to do any work. The biggest challenge we face while collecting data from Facebook and twitter was many similar news was around with different headlines so we had to filter them out one by one. It is necessary to collect proper and relevant news headlines, whether fake or not for a well working dataset.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

This section lays out the basis for the investigation's research techniques. We also discuss the methods used to break down the data. Finally, the implementation requirements that were met in the project are discussed.

3.2 Research Subject and Instrumentation

Our analysis focuses on detecting fake and real news. We're doing it with a variety of machine learning algorithms. "A comparative analysis on detecting Bangla fake news on social media using machine learning algorithms" is our research subject. By browsing news portals, we gathered all of the data from Facebook and Twitter. We used laptops and mobile devices, as well as our own social media accounts, as instruments for our study. Following the data collection, several preprocessing techniques were used to generalize the data. After that, we prepared the preprocessed data for our research. The prepared data was then fed into machine learning (ML) algorithms in order to predict our result.

3.3 Data Collection Procedure

One of the most critical aspects of our study is data collection. Data collection, also known as information gathering, is the process of assembling and analyzing data on factors of interest in a structured manner that permits one to respond to specified research questions, test assumptions, and determine outcomes. All sectors of study, including in presence and sociologies, humanities, industry, and so on, depend on the information gathering aspect of research. Though methods vary by discipline, the importance on confirming correct and equal accumulation doesn't have a noticeable change. It was one of the most difficult challenges we have faced. There are several methods for gathering information and data,

but we chose to collect our data from Facebook and twitter. Although google also helped to navigate the occurrence time of a certain news.

3.4 Statistical Analysis

1000 of each fake and real dataset was collected. Here is a visualization of both datasets.

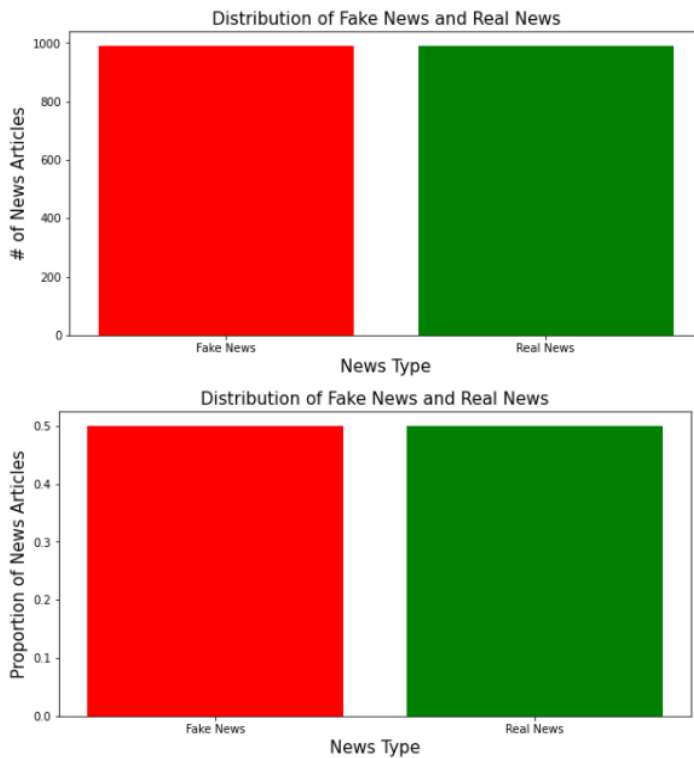


Fig 3.4.1: Statistics of datasets

3.5 Algorithms and procedure:

Step 1:

We imported the datasets into colab by accessing the google drive. The datasets named Real datasets and fake datasets respectively we have created a manual testing dataset for testing manually consisting last 10 rows from each dataset that shortens the datasets by 10 rows.

- We have used four machine learning algorithms in our research.
 1. Logistic regression

- 2. Decision tree classification
- 3. Gradient Boosting Classifier
- 4. Random Forest Classifier

Step 2:

We dropped the unnecessary column from the merged dataset.

```
[ ] df = df_merge.drop(["headline"], axis = 1)
```

Fig 3.5.1: dropped column

Step 3:

After shuffling the data frame randomly we created a function to **convert the text in lowercase, remove the extra space, special characters, URL and links. After we defined dependent and independent variables as x and y, we split the dataset into training dataset and testing dataset.**

Step 4:

For text processing we have used tfidfvectorizer. All described broadly below:

Tfidfvectorizer:

The TF-IDF is a subtask of information retrieval and extraction that seeks to convey the value of a word to a text that is part of a corpus (a collection of documents). Few search engines use it to get improved results that are more important to a particular enquiry. In this section, we'll go over what TF-IDF is and how it works and explain the math behind this one, and then show how to use the SK-Learn library to implement it in Python.

Term Frequency (TF): The Term Frequency of a word is the amount of times it comes in a sentence. If a word appears more often than others, a higher value indicates that the text is a good fit when the term is part of the search words.

IDF (Inverse Document Frequency): Words that appear frequently in one document but not in others may be irrelevant. The IDF is a metric for determining how important a word is across the entire corpus.

The TfidfVectorizer transforms a pile of raw documents into a TF-IDF function matrix.

This is done by counting the amount of times a word happens to be in a document as well as the number of times the identical word occurs in other documents in the corpus.

The reasoning for this as follows:

- A term that pops up often in a news has greater significance for that news, implying that the document is more likely to be directing towards that particular word.

- A word that appears regularly in more news will make it difficult to find the right news, the word is appropriate for every news or no news at all. It won't help us strain out a single or a tiny portion of news from the entire collection in either case.
- As a result, in our dataset TF-IDF is indeed a result that is applied to each and all words in each text. And the TF-IDF value for each word increases with each occurrence in a news, but progressively decreases with each arrival in other papers or documents. The math for it is in the following section:

Then let's have a look at the TF-IDF numerical measure's basic formula. Let us start by defining few terms:

N, (the number of documents we have in our dataset)

d, (a given document from our dataset)

D, (the collection of all documents)

w, (given word in a document)

now the term frequency equation is:

$$tf(w, d) = \log(1 + f(w, d))$$

The frequency of the word w in document d is f (w,d).

Now the inverse frequency:

$$idf(w, D) = \log\left(\frac{N}{f(w, D)}\right)$$

Now the final step tfidf computing:

$$tfidf(w, d, D) = tf(w, d) * idf(w, D)$$

Step 5:

The algorithms mentioned below were used and described briefly:

3.5.1 Logistic regression:

The representation of logistic regression, that of linear regression, is an equation.

In predicting an output value, input values (x) are linearly combined by means of weights or coefficient values (abbreviated as Beta in Greek) (y). The output result is a binary value (0 or 1), not a numeric number, which is a key contrast from linear regression.

Here is the equation:

$$y = e^{(b_0 + b_1 * x)} / (1 + e^{(b_0 + b_1 * x)})$$

Where y is the expected output, b_0 is the bias or intercept expression, and b_1 is the single input value coefficient (x). The b coefficient (a constant real value) for each column throughout the input data must be learned from the training data.

Logistic regression has many advantages: it can model probabilities (unlike decision trees and SVMs), features can be dependent (unlike Nave Bayes), and it is simple to update the model with new data.

For the library we used sklearn. As the parameters in LogisticRegression function, we assigned 'l2' (default) in penalty, it specifies the norm that is used in penalization.

Other parameters dual, fit_intercept, C, intercept_scaling, ratio, max_iter, multi_class, n_jobs, tol, verbose, random_state, solver, warm_state are all set to their default values.

The LR.score shows 0.89.

Here is the classification report:

	precision	recall	f1-score	support
0	0.91	0.87	0.89	241
1	0.88	0.92	0.90	254
accuracy			0.90	495
macro avg	0.90	0.90	0.90	495
weighted avg	0.90	0.90	0.90	495

Table: 3.5.1: Logistic regression classification report

3.5.2 Decision tree classification:

In a circumstance of a tree structure, a decision tree builds classification or regression models. It gradually breakdowns a dataset into some reduced subsets whereas developing a related decision tree also. It's like a flowchart, is a group of decision nodes with a tree graph structure. We start from the beginning when making a decision. The input (news headline information) is checked at each decision node, and we proceed along a branch to the next node based on the outcome. We can use the formula to measure a node's entropy:

$$E(S) = \sum_{i=1}^c -p_i \log_2(p_i)$$

We used sklearn library, in the DecisionTreeClassifier function, in ccp_alpha we assigned 0.0 as non-negative float as default value. It used to prone minimal cost-complexity. In criterion parameter, used to quantify the split of quality the value assigned is default 'mse' means mean squared error.

Other parameters including class_weight, max_depth, max_features, max_leaf_nodes, min_impurity_decrease, min_impurity_split, min_samples_leaf and all the other parameters were set to default.

The classification report is given below:

	precision	recall	f1-score	support
0	0.82	0.81	0.81	241
1	0.82	0.83	0.82	254
accuracy			0.82	495
macro avg	0.82	0.82	0.82	495
weighted avg	0.82	0.82	0.82	495

Table: 3.5.2: Decision tree classification report

3.5.3 Gradient Boosting Classifier:

To improve accuracy, Machine learning algorithms need more than just suitable models and making predictions. For enhanced work, most persuasive models in the business or in competitions have been using ensemble techniques or feature engineering. In comparison to Feature Engineering, Ensemble techniques have grown admiration for the easiness of the procedure. When applied with innovative machine learning algorithms, a variety of ensemble approaches have been shown to improve accuracy. Gradient Boosting is one such process. Its a collection of machine learning algorithms that incorporate a number of weak learning models to make a better predictor model.

With large and complex data, its an effective technique to predict and to get a result. It is based on the assumption that when the maximum suited next model is consolidated with previous models, the overall prediction error is lessened. It trained model in a more sequential manner.

We used sklearn library, in the function we assigned 0.0 as float value in the parameter of ccp_alpha. Its for the pruning of minimal cost-complexity. The subtree with the highest cost complexity (less than ccp alpha) will be picked. The other parameters were set to default also.

The classification report is below:

	precision	recall	f1-score	support
0	0.92	0.90	0.91	250
1	0.90	0.92	0.91	245
accuracy			0.91	495
macro avg	0.91	0.91	0.91	495
weighted avg	0.91	0.91	0.91	495

Table: 3.5.3: Gradient Boosting classification report

3.5.4 Random Forest Classifier:

Random Forest is a vastly used machine learning algorithm, it uses supervised learning method. It could be used for equally classification and regression problems in machine

learning, It is based on collaborative learning, A method of combining multiple classifiers to resolve a compound problem and advance the models accuracy.

Here is the diagram of how random forest classifier works:

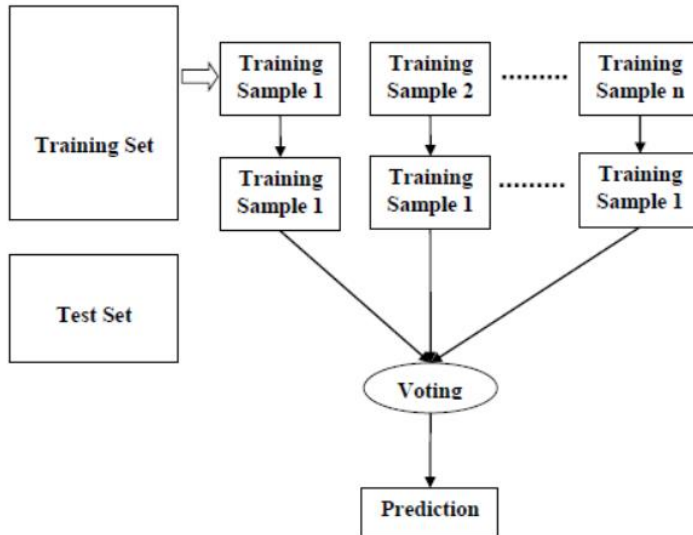


Fig 3.5.4: Random forest classifier diagram

We used sklearn library. In the function RandomForestClassifier, the bootstrap is set to “True”, if false, each tree will be used to build each tree. For two sample splits, the min_sample_split is set to 2. Other parameters are set to their default values.

Here is the classification report:

	precision	recall	f1-score	support
0	0.91	0.91	0.91	241
1	0.92	0.91	0.92	254
accuracy			0.91	495
macro avg	0.91	0.91	0.91	495
weighted avg	0.91	0.91	0.91	495

Table: 3.5.4: Random Forest classification report

3.6 Proposed framework:

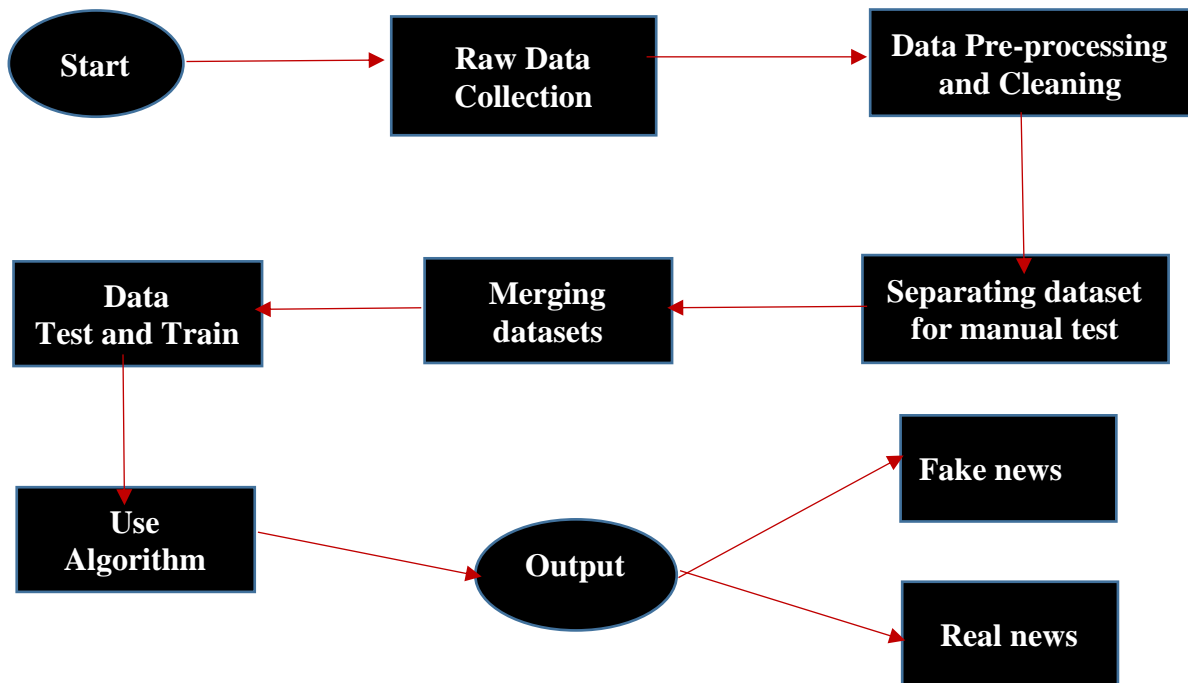


Figure 3.6.1: Framework.

For this study, the reliability of fake news is being tested. Early detection of fake news could prevent a lot of bad things from happening. First and foremost, data was gathered through the internet. Data preprocessing methods were used to prepare and clean the collected data. The data was then generalized. Finally, when it is given to the algorithm for prediction, an output is provided by the algorithm.

Our expected performance will be estimated using this methodology. When it comes to algorithms, there have been four algorithms used to predict fake news. Logistic regression, decision tree classification, gradient boosting classifier, and random forest classifier are the algorithms used. We use four algorithms to distinguish the accuracy rates of these four, and then a confusion matrix to evaluate the algorithm's performance. We'll need to import some library functions into our IDE to complete all of these tasks. For this fake news identification, we used Google Colab as our main working platform. The libraries that were used are listed below:

Used libraries:

A screenshot of a code editor interface. On the left, there is a vertical sidebar with a '<>' icon at the top and a '□' icon below it. The main area of the editor has a dark background and displays Python code for importing various libraries. The code is as follows:

```
[ ] import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import re
import string
```

Figure 3.6.2: Used Library

We have used pandas, numpy library in our project. Pandas is used for data frame. Sklearn.model_selection is used for importing the models or algorithm which we want to use and also used for importing train_test_split function for train and test the dataset.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

4.1 Introduction

Out of the four algorithms (Logistic regression, Decision tree classification, Gradient Boosting Classifier, Random Forest Classifier) Random forest classifier seems to be able to detect more accurately. Our model was able to detect almost every input from manual test dataset whether the input news is fake or not. Early detection of fake news could stop an individual from being scammed or miss directed.

4.2 Experimental Results

Accuracy of algorithms:

Algorithm	Accuracy
Logistic regression	0.896969696969697
Decision tree classification	0.8181818181818182
Gradient Boosting Classifier	0.8848484848484849
Random Forest Classifier	0.9131313131313131

Table: 4.2.1: Accuracy table

Here we can see among the four algorithms, Random forest classifier has an accuracy of 91% where logistic regression came out with 89%, Decision tree classifier came out with an accuracy of 81% and Gradient boosting classifier came out with an 88% of accuracy.

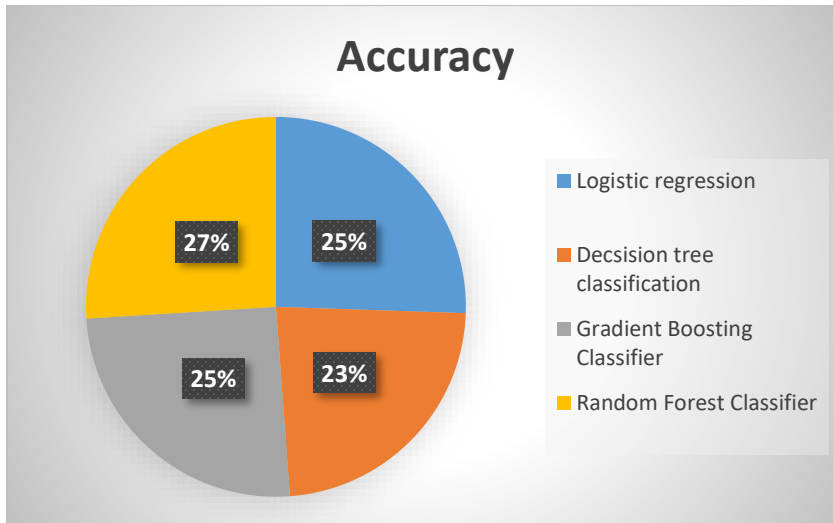


Fig 4.2.2: Pie chart

Confusion Matrix:

Confusion matrix is a counter that is regularly used to depict the exhibition of an order model on a lot of test information for which the genuine qualities are recognized. It permits the perception of the exhibition of a calculation. Confusion matrix have been found for our model.

Logistic regression:

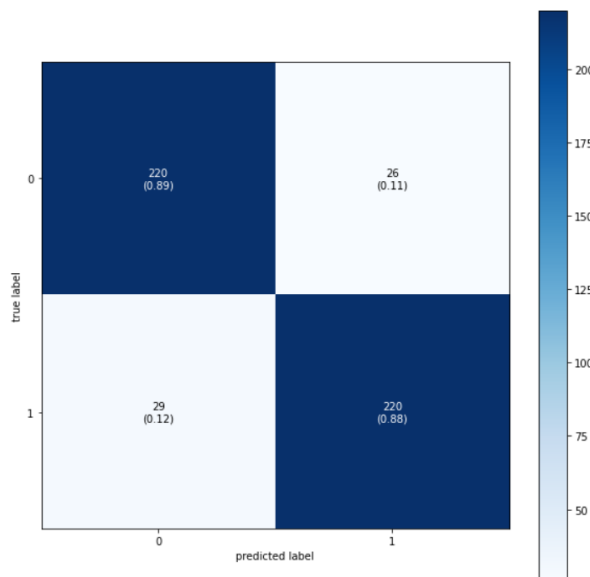


Fig 4.2.3: Confusion matrix of Logistic regression

Decision tree classification:

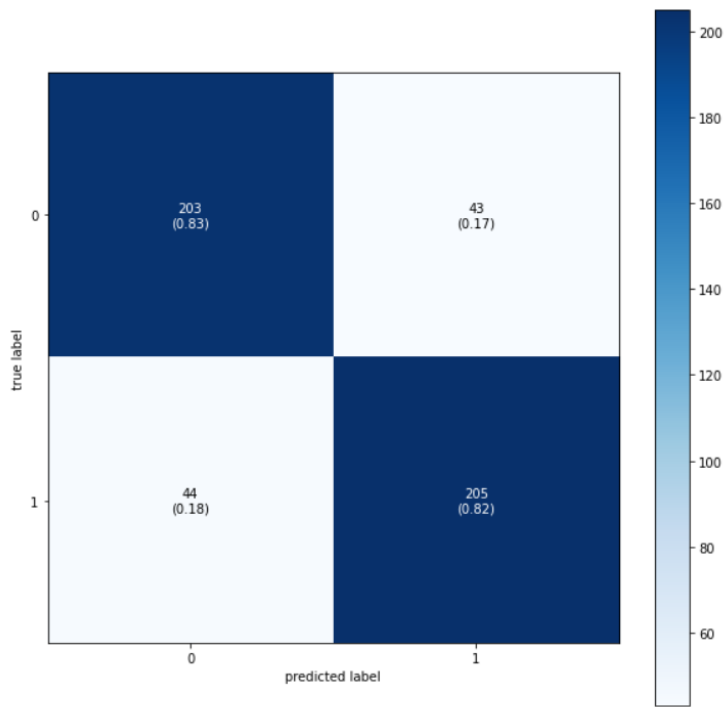


Fig 4.2.4: Confusion matrix of Decision tree classification

Gradient Boosting Classifier:

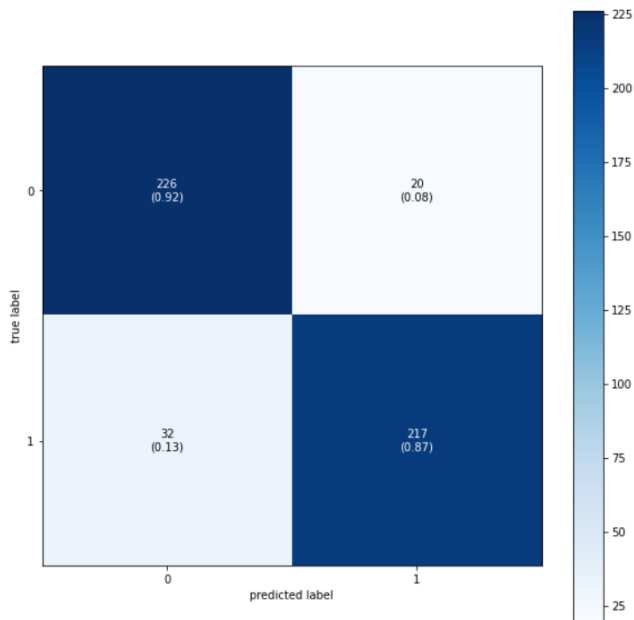


Fig 4.2.5: Confusion matrix of Gradient Boosting Classifier

Random Forest Classifier:

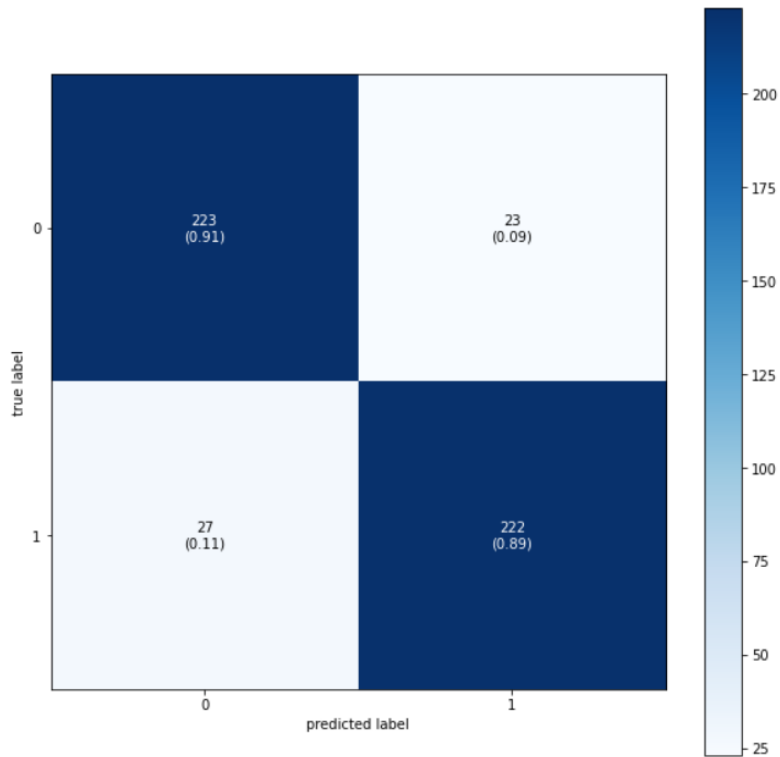


Fig 4.2.6: Confusion matrix of Random Forest Classifier

4.3 Summary

In this work we can see the random forest algorithm has an accuracy of 91% and more accurate than any other algorithms. Logistic regression was also close but after the work we can draw a conclusion that Logistic regression have a good accuracy than other three algorithms.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Summary of the Study

Fake news is one of the most well-known issues in our everyday lives, especially online. It has had a major impact in our nation for many years as well. Many people have been harmed, deceived, or misdirected in various ways. As a result, we must teach people to be more cautious and mindful of whether the news they are reading is true or not in order to avoid being duped. When we first considered this study, we wondered if we could get people to be more cautious about believing news right away. As a result, a model with an easy-to-use input option has been created, allowing people to simply type in the headline of the news to determine whether it is accurate or not, using a different machine learning algorithm. Over our collected data and dataset, we used logistic regression, Decision tree classification, Gradient boosting classifier, and Random Forest Classifier to detect fake news. The most efficient was the Random forest classifier. The most difficult aspect of our study was gathering data for dataset.

5.2 Conclusions

Misleading content, like fake news and fake review of goods, has turned out to be a more risky prospect for internet consumers in latest years. Fake news and fake reviews of many products have harmed both humanity and businesses. The use of employed writers to create fake reviews in order to increase transactions is also on the rise. According to recent figures and studies, 62% of adults in the USA receiving their news via social and online media. The majority of prominent fake news stories received more Facebook shares than any popular mainstream social media.

5.3 Recommendations

Machine Learning (ML) is the logical learning of equations and plausible models that used by computer frameworks to carry out an exact task deprived of using categorical instructions, trusting instead on examples and guesswork. Artificial intelligence is regarded as a subset of it. Machine Learning calculations construct a numerical model based on test data, which is referred to as training data, so that priorities or options can be set without being explicitly changed in order to carry out the work To improve our model for detecting fake news and demonstrating which is stronger in prediction over our data set, we used four different machine learning algorithms these are Logistic regression, Decision tree classification, Gradient boosting classifier, and Random forest classifier are the four algorithms. While conducting research on our dataset, we discovered that these various algorithms have a nearby accuracy score. While algorithm accuracy varies from dataset to dataset, the accuracy rates for the algorithms used in our dataset are almost identical. When we increase or decrease our test value the accuracy of Random Forest Classifier is higher than other algorithms. So we came to a conclusion to recommend to use or try Random Forest Classifier in detecting fake news research work.

5.4 Further Study

In this work a model has been developed for detecting fake or real news using different machine learning algorithms but in the future the main focus will be to use use machine learning algorithms to predict the pattern and how it's designed to make a fake news looks like a real one so that it could be filtered out before it gets into internet. I consider it will be also more useful for people to find news that they read is trustable or not. Also we will collect more data to make dataset big to get more accurate results.

References

- [1] A. Bessi, M. Coletto, G. A. Davidescu, A. Scala, G. Caldarelli, and W. Quattrociocchi (2015), “Science vs Conspiracy: Collective Narratives in the Age of Misinformation”, PLOS ONE, vol. 10(2), pp. e0118093.
- [2] H. Allcott and M. Gentzkow (2017), “Social Media and Fake News in the 2016 Election”, *Journal of Economic Perspectives*, vol. 31(2), pp. 211–236.
- [3] A. Guess, J. Nagler, and J. Tucker (2019), “Less than you think: Prevalence and predictors of fake news dissemination on Facebook”, *Science Advances*, vol. 5(1), pp.
- [4] M. Balmas (2012), “When Fake News Becomes Real: Combined Exposure to Multiple News Sources and Political Attitudes of Inefficacy, Alienation, and Cynicism”, *Communication Research*, vol. 41(3), pp. 430-454.
- [5] S. Jang, and K. K. Joon (2018). “Third person effects of fake news: Fake news regulation and media literacy interventions”, *Computers in Human Behavior*, vol. 80, pp. 295-302.
- [6] H. Karimi, P. C. Roy, S. S. Sadiya, and J. Tang (2018), “Multi-Source Multi-Class Fake News Detection”, *Proceedings of the 27th International Conference on Computational Linguistics*, New Mexico, USA, pp. 1546– 1557.
- [7] A. Bondielli and F. Marcelloni, “A survey on fake news and rumour detection techniques,” *Information Sciences*, vol. 497, p. 3855, 2019.
- [8] R. K. Nielsen, R. Fletcher, N. Newman, J. S. Brennen, and P. N. Howard, “Navigating the ‘infodemic’: How people in six countries access and rate news and information about coronavirus,” Apr 2020.
- [9] U. News. During this coronavirus pandemic, fake news is putting lives at risk: Unesco. [Online]. Available: <https://news.un.org/en/story/2020/04/1061592>
- [10] <https://www.thedailystar.net/city/news/they-tried-destroy-communal-harmony-1806874>
- [11] R. Mihalcea, and C. Strapparava. ”The lie detector: Explorations in the automatic recognition of deceptive language.” *Proceedings of the ACL-IJCNLP Conference Short Papers*. Association for Computational Linguistics, 2009.
- [12] S. Feng, R. Banerjee, and C. Yejin, ”Syntactic stylometry for deception detection.” *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 2012.
- [13] H. Zhen, et al. ”Deceptive review spam detection via exploiting task relatedness and unlabeled data.” *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016
- [14] A. Borodin, G. O. Roberts, J. S. Rosenthal, and P. Tsaparas. Link analysis ranking: algorithms, theory, and experiments. *TOIT*, 5(1):231–297, 2005.
- [15] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating web spam with trustrank. In *VLDB*, 2004.
- [16] G. L. Ciampaglia, P. S., L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini. Computational fact checking from knowledge networks. *PLOS ONE*, 10(6):1–13, 06 2015.

- [17] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy, S. Sun, and W. Zhang. From data fusion to knowledge fusion. In VLDB, 2014.
- [18] X. L. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang. Knowledge-based trust: Estimating the trustworthiness of web sources. In VLDB, 2015.
- [19] H. Xiao, J. Gao, Q. Li, F. Ma, L. Su, Y. Feng., and A. Zhang. Towards confidence in the truth: A bootstrapping based truth discovery approach. In KDD, 2016.
- [20] S. Kumar, R. West, and J. Leskovec. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In WWW, 2016.
- [21] C. Lumezanu, N. Feamster, and H. Klein. # bias: Measuring the tweeting behavior of propagandists. In ICWSM, 2012
- [22] P. A. Lewis and G. S. Shedler. Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics*, 26(3):403–413, 1979.
- [23] B. T. Adler and L. De Alfaro. A content-driven reputation system for the wikipedia. In WWW, 2007
- [24] Z. Zhao, P. Resnick, and Q. Mei. Enquiring minds: Early detection of rumors in social media from enquiry posts. In WWW, 2015.
- [25] M. Tanushree, G. Wright, and E. Gilbert. Parsimonious language model of social media credibility across disparate events. In CSCW, 2017.
- [26] W. Wei and X. Wan. Learning to identify ambiguous and misleading news headlines. In IJCAI, 2017.
- [27] S. Volkova, K. Shaffer, J. Y. Jang, and N. Hodas. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on twitter. In ACL, 2017.
- [28] A. Gupta, P. Kumaraguru, C. Castillo, and P. Meier. Tweetcred: Real-time credibility assessment of content on twitter. In SocInfo, 2014.
- [29] N. Ruchansky, S. Seo, and Y. Liu. Csi: A hybrid deep model for fake news detection. In CIKM, 2017
- [30] J. Pasternack and D. Roth. Latent credibility analysis. In WWW, 2013.
- [31] X. Yin and W. Tan. Semi-supervised truth discovery. In WWW, 2011
- [32] B. Zhao and J. Han. A probabilistic model for estimating real-valued truth from conflicting sources. In QDB, 2012.
- [33] B. Zhao, B. I. Rubinstein, J. Gemmell, and J. Han. A bayesian approach to discovering truth from conflicting sources for data integration. In VLDB, 2012.
- [34] B. Tabibian, I. Valera, M. Farajtabar, L. Song, B. Schoelkopf, and M. Gomez-Rodriguez. Distilling information reliability and source trustworthiness from digital traces. In WWW, 2017.
- [35] M. Lukasik, P. K. Srijith, D. Vu, K. Bontcheva, A. Zubiaga, and T. Cohn. Hawkes processes for continuous time sequence classification: an application to rumour stance classification in twitter. In ACL, 2016
- [36] S. Badaskar, S. Agarwal, and S. Arora (2008), “Identifying Real or Fake Articles: Towards better Language Modeling”, *IJCNLP*, pp. 817–822

- [37] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro (2017), “Some Like it Hoax: Automated Fake News Detection in Social Networks”, Proceedings of the Second Workshop on Data Science for Social Good, Skopje, Macedonia, vol. 1960.
- [38] E. Tacchini, G. Ballarin, M. L. Della Vedova, S. Moret, and L. de Alfaro (2017), “Some Like it Hoax: Automated Fake News Detection in Social Networks”, Proceedings of the Second Workshop on Data Science for Social Good, Skopje, Macedonia, vol. 1960.
- [39] H. Karimi, P. C. Roy, S. S. Sadiya, and J. Tang (2018), “Multi-Source Multi-Class Fake News Detection”, Proceedings of the 27th International Conference on Computational Linguistics, New Mexico, USA, pp. 1546– 1557.
- [40] K. Shu, D. Mahudeswaran, and H. Liu (2018), “FakeNewsTracker: a tool for fake news collection, detection, and visualization”, Computational and Mathematical Organization Theory, vol. 25(1), pp. 60-71

APPENDICES

Appendix

▾ Importing required library

```
[ ] import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
from sklearn.metrics import classification_report
import re
import string
```

Inserting Real and fake dataset

```
[ ] from google.colab import drive

drive.mount('/content/gdrive')
```

Mounted at /content/gdrive

```
[ ] import pandas as pd
df_real = pd.read_excel('/content/gdrive/MyDrive/Colab Notebooks/Real.xlsx')
df_fake = pd.read_excel('/content/gdrive/MyDrive/Colab Notebooks/Fake.xlsx')
```

```
[ ] df_real.head(5)
```

	headline	article	label
0	গ্রীসের সঙ্গে দ্বন্দ্ব চরমে, সাইপ্রাসে আরো সেন...	আন্তর্জাতিক ডেস্ক আরটিএনএন আঙ্কারা: তুরস্কের প...	1
1	সিরিয়ায় রুশ বিমান ভূপাতিত, ইসরাইলকে দুষলো রাশিয়া	আন্তর্জাতিক ডেস্ক আরটিএনএন বৈরুত: সিরিয়া দুর্ঘ...	1
2	রোহিঙ্গাদের ওপর সামরিক বাহিনী নৃশংসতার 'পরিমাপ...	আন্তর্জাতিক ডেস্ক আরটিএনএন জেনেভা: জাতিসংঘের ত...	1
3	ইন্ডিগ্রাম কিনতে ৪ হাজার কোটি টাকার প্রকল্প একনেকে ...	নিজস্ব প্রতিবেদক আরটিএনএন ঢাকা: নির্বাচন সামনে...	1
4	ভারত-বাংলাদেশের সম্পর্ক অনন্য উচ্চতায় উন্নীত হ...	নিজস্ব প্রতিবেদক আরটিএনএন ঢাকা: প্রধানমন্ত্রী ...	1

```
[ ] df_fake.head(5)
```

	headline	article	label
0	মুরগির হামলায় শেয়াল নিহত	বাংলায় একটা প্রবাদ আছে, শেয়ালের কাছে মুরগী বর্...	0
1	বিটিভিতে যেবার আমি ইন্টারভিউ দিতে গেলাম	BTV থেকে লোকজন আসছে, ইন্টারভিউ নিবে।চারজনের টি...	0
2	বিদেশ থেকে উন্নতমানের বিরোধীদল আমদানি করার পরা...	অদ্ভুত বিরোধীদলহীনতায় ভুগছে সরকার। এ এক অন্যরক...	0
3	অবসর নেয়ার ঘোষণা দিলেন মেসি !	রাশিয়া বিশ্বকাপ নকআউট পর্বে ফ্রান্সের সাথে ৪-৩...	0
4	মাদারফলকার নহে, ব্রাদারফলকার: সাকা দৈনিক মতি...	নিজস্ব প্রতিবেদক মাদারফলকার নহে, আমি ব্রাদারফলকা...	0

```
[ ] df_fake.shape, df_real.shape
```

```
((999, 3), (999, 3))
```

Removing last 10 rows from both the dataset, for manual testing

```
[ ] df_fake_manual_testing = df_fake.tail(10)
    for i in range(998,989,-1):
        df_fake.drop([i], axis = 0, inplace = True)
df_real_manual_testing = df_real.tail(10)
for i in range(998,989,-1):
    df_real.drop([i], axis = 0, inplace = True)
```

```
[ ] df_fake.shape, df_real.shape
```

```
((990, 3), (990, 3))
```

```
[ ] plt.figure(figsize=(10, 5))
    plt.bar('Fake News', len(df_fake), color='red')
    plt.bar('Real News', len(df_real), color='green')
    plt.title('Distribution of Fake News and Real News', size=15)
    plt.xlabel('News Type', size=15)
    plt.ylabel('# of News Articles', size=15)
```

```
total_len = len(df_fake) + len(df_real)
plt.figure(figsize=(10, 5))
plt.bar('Fake News', len(df_fake) / total_len, color='red')
plt.bar('Real News', len(df_real) / total_len, color='green')
plt.title('Distribution of Fake News and Real News', size=15)
plt.xlabel('News Type', size=15)
plt.ylabel('Proportion of News Articles', size=15)
```

```
Text(0, 0.5, 'Proportion of News Articles')
```

Merging the manual testing dataframe in single dataset and saving it in a excel file

```
[ ] df_fake_manual_testing["class"] = 0
df_real_manual_testing["class"] = 1
```

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
 """Entry point for launching an IPython kernel.

```
/usr/local/lib/python3.7/dist-packages/ipykernel_launcher.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead
```

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy

```
[ ] df_fake_manual_testing.head(10)
```

	headline	article	label	class
989	মনোনয়ন নেবো না, মনোনয়ন দেবো : যারা মনোনয়ন পানন...	প্রকাশিত হয়েছে মনোনয়নের রেজাল্ট। অনেকেই আসন্ন ...	0	0
990	যাত্রা শুরু করলো ওসমান'স ব্রাইডাল মেকওভার সেলুন	এসে গেছে শীতকাল, মানে বিয়ের মৌসুম। এই সময়টিতে ...	0	0
991	Zee বাংলা বন্ধুর সংবাদ শুনে সীমান্ত এলাকায় টিভ...	হঠাৎ করেই সারাদেশে জি বাংলা বন্ধ হয়ে যাওয়াতে স...	0	0
992	যে ১০টি প্রশ্নের উত্তর সম্পর্কে দেবী সিনেমার র...	মুক্তি পেয়েছে ছমাযুন আহমেদের 'দেবী' উপন্যাস অ...	0	0
993	বই পড়ার সময় আমি অন্য কিছু করি না কিন্তু... - B...	"মদ খাওয়ার সময় আমি কোন রিস্ক নেই না" গল্প অবলম...	0	0
994	প্রেম করলে বাড়বে গুজনা	গুজন বাড়ার সঙ্গে নাকি প্রেমের এক অদ্ভুত সম্পর্...	0	0
995	ঈদের ছুটিতে রাস্তাঘাট বেশি ফাঁকা দেখে ডিপ্রেশন...	ঈদের ছুটি মানেই রাস্তাঘাট ফাঁকা, মানুষের ভীড় আ...	0	0
996	ঢাকার যানজটকে হাসিমুখে মেনে নিতে গঠিত হলো 'হাস...	ঢাকার রাস্তার জ্যাম ঢাকার গর্ব, গিনেসবুকে নাম ...	0	0
997	স্বঘোষিত নিষেধাজ্ঞার কারণে এবার নৌকায় ভোট দিতে...	আসন্ন জাতীয় নির্বাচনে নৌকা মার্কায় ভোট দিতে পা...	0	0
998	অবাধ সূষ্ঠ নির্বাচন হলে বিএনপির সবার জামানত বা...	নির্বাচনের সময় নেতাকর্মীরা সাধারণত নিজেদের দল ...	0	0

✓ 30s completed at 1:15 AM

Merging the main fake and real dataframe

```
[ ] df_merge = pd.concat([df_fake, df_real], axis = 0 )
df_merge.head(10)
```

	headline	article	label
0	মুরগির হামলায় শেয়াল নিহত	বাংলায় একটা প্রবাদ আছে, শেয়ালের কাছে মুরগী বর্...	0
1	বিটিভিতে য়েবার আমি ইন্টারভিউ দিতে গেলাম	BTV থেকে লোকজন আসছে, ইন্টারভিউ নিবে।চারজনের টি...	0
2	বিদেশ থেকে উন্নতমানের বিরোধীদল আমদানি করার পরা...	অদভূত বিরোধীদলহীনতায় ভুগছে সরকার। এ এক অন্যরক...	0
3	অবসর নেয়ার ঘোষণা দিলেন মেসি!	রাশিয়া বিশ্বকাপ নকআউট পর্বে ফ্রান্সের সাথে ৪-৩...	0
4	মাদারফাকার নহে, ব্রাদারফাকার: সাকা দৈনিক মতি...	নিজস্ব মতিবেদক মাদারফাকার নহে, আমি ব্রাদারফাকা...	0
5	বিয়ের পিড়িতে বসছেন মিয়া খলিফা। ছেলে কুমিল্লার	বিয়ের সানাই বাজতে চলেছে শীঘ্রই। সব জল্পনা কল...	0
6	জুম্মার নামাজে সবচেয়ে বেশি মসজিদে যায় নোয়াখ...	এক গবেষণা থেকে জানা গেছে, বাংলাদেশের অন্যান্য ...	0
7	প্রধানমন্ত্রীর প্রশ্ন: আমনে আমান্তে বড় দেশপ্রে...	নিজস্ব মতিবেদকতেল-গ্যাস-খনিজ সম্পদ ও বিদ্যুৎ-ব...	0
8	জানেন শিব ঠাকুরের বাবা কে? জেনে নিন তাহলে...-	তেত্রিশ কোটি দেবতার মধ্যে এক-একজন এক-এক বেশে এ...	0
9	মেডামের দুয়ায় সমস্যা আছে: মিছবাউল দৈনিক মতিকণ্ঠ	ক্রীড়া মতিবেদকচলমান বিশ্বকাপ কুকেটে পাকিস্তানে...	0

```
[ ] df_merge.columns
Index(['headline', 'article', 'label'], dtype='object')
```

```
[ ] df_merge.columns
Index(['headline', 'article', 'label'], dtype='object')
```

I need "article" and "label" column so dropping other columns

```
[ ] df = df_merge.drop(["headline"], axis = 1)
```

```
[ ] df.isnull().sum()
```

```
article    0
label      0
dtype: int64
```

Randomly shuffling the dataframe

```
[ ] df = df.sample(frac = 1)
```

```
[ ] df.head()
```

	article	label
779	কাটমন্ডু মতিনিধি৪০ বতসর বয়সী চীনা নারী পর্বতার...	0
645	জ্যেষ্ঠ প্রতিবেদক : বিএনপি নির্বাচনে গেলে আওয়া...	1
60	দুবাই আন্তর্জাতিক ক্রিকেট স্টেডিয়ামে বুধবার প...	1
89	'ভালো নেই আফজাল শরীফ, সহায়তা চাইলেন প্রধানমন্ত...	1

Randomly shuffling the dataframe

```
[ ] df = df.sample(frac = 1)
```

```
[ ] df.head()
```

	article	label
779	কাটমন্ডু মতিনিধি৪০ বতসর বয়সী চীনা নারী পর্বতার...	0
645	জ্যেষ্ঠ প্রতিবেদক : বিএনপি নির্বাচনে গেলে আওয়া...	1
60	দুবাই আন্তর্জাতিক ক্রিকেট স্টেডিয়ামে বুধবার প...	1
89	'ভালো নেই আফজাল শরীফ, সহায়তা চাইলেন প্রধানমন্ত...	1
388	আজ সকালে ফেসবুক কর্তৃপক্ষ আমাদের নিউজ ডেস্কে ফ...	0

```
[ ] df.reset_index(inplace = True)
df.drop(["index"], axis = 1, inplace = True)
```

```
[ ] df.columns
```

```
Index(['article', 'label'], dtype='object')
```

```
[ ] df.head()
```

	article	label
0	কাটমন্ডু মতিনিধি৪০ বতসর বয়সী চীনা নারী পর্বতার...	0

Creating a function to convert the text in lowercase, remove the extra space, special chr., ulr and links

```
[ ] def wordopt(article):
    article = article.lower()
    article = re.sub('\[.*?\]', '', article)
    article = re.sub("\W", " ", article)
    article = re.sub('https?://\S+|www\.\S+', '', article)
    article = re.sub('<.*?>+', '', article)
    article = re.sub('%s' % re.escape(string.punctuation), '', article)
    article = re.sub('\n', '', article)
    article = re.sub('\w*\d\w*', '', article)
    return article
```

```
[ ] df["article"] = df["article"].apply(wordopt)
```

Defining dependent and independent variable as x and y

```
[ ] x = df["article"]
    y = df["label"]
```

Splitting the dataset into training set and testing set.

```
[ ] x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

Convert text to vectors

```
[ ] from sklearn.feature_extraction.text import TfidfVectorizer
```

```
[ ] vectorization = TfidfVectorizer()
    xv_train = vectorization.fit_transform(x_train)
    xv_test = vectorization.transform(x_test)
```

1. Logistic regression

```
[ ] from sklearn.linear_model import LogisticRegression

[ ] LR = LogisticRegression()
LR.fit(xv_train,y_train)

LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
intercept_scaling=1, l1_ratio=None, max_iter=100,
multi_class='auto', n_jobs=None, penalty='l2',
random_state=None, solver='lbfgs', tol=0.0001, verbose=0,
warm_start=False)

[ ] pred_lr=LR.predict(xv_test)

[ ] LR.score(xv_test, y_test)

0.896969696969697

[ ] LR.score(xv_test, y_test)

0.896969696969697

[ ] print(classification_report(y_test, pred_lr))

              precision    recall  f1-score   support

     0           0.91     0.87     0.89         241
     1           0.88     0.92     0.90         254

 accuracy          0.90
 macro avg          0.90
weighted avg          0.90
```

20% completed at 1:15 AM

2. Decision tree classification

```
[ ] from sklearn.tree import DecisionTreeClassifier

[ ] DT = DecisionTreeClassifier()
DT.fit(xv_train, y_train)

DecisionTreeClassifier(ccp_alpha=0.0, class_weight=None, criterion='gini',
max_depth=None, max_features=None, max_leaf_nodes=None,
min_impurity_decrease=0.0, min_impurity_split=None,
min_samples_leaf=1, min_samples_split=2,
min_weight_fraction_leaf=0.0, presort='deprecated',
random_state=None, splitter='best')

[ ] pred_dt = DT.predict(xv_test)

[ ] DT.score(xv_test, y_test)

0.8181818181818182

[ ] print(classification_report(y_test, pred_dt))

              precision    recall  f1-score   support

     0           0.82     0.81     0.81         241
     1           0.82     0.83     0.82         254

 accuracy          0.82
 macro avg          0.82
weighted avg          0.82

[ ] from sklearn.metrics import classification_report, confusion_matrix
from mlxtend.plotting import plot_confusion_matrix
```

▼ 3. Gradient Boosting Classifier

```
[ ] from sklearn.ensemble import GradientBoostingClassifier
```

```
[ ] GBC = GradientBoostingClassifier(random_state=0)
GBC.fit(xv_train, y_train)
```

```
GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', init=None,
                           learning_rate=0.1, loss='deviance', max_depth=3,
                           max_features=None, max_leaf_nodes=None,
                           min_impurity_decrease=0.0, min_impurity_split=None,
                           min_samples_leaf=1, min_samples_split=2,
                           min_weight_fraction_leaf=0.0, n_estimators=100,
                           n_iter_no_change=None, presort='deprecated',
                           random_state=0, subsample=1.0, tol=0.0001,
                           validation_fraction=0.1, verbose=0,
                           warm_start=False)
```

```
[ ] GradientBoostingClassifier(random_state=0)
```

```
GradientBoostingClassifier(ccp_alpha=0.0, criterion='friedman_mse', init=None,
                           learning_rate=0.1, loss='deviance', max_depth=3,
                           max_features=None, max_leaf_nodes=None,
                           min_impurity_decrease=0.0, min_impurity_split=None,
                           min_samples_leaf=1, min_samples_split=2,
                           min_weight_fraction_leaf=0.0, n_estimators=100,
                           n_iter_no_change=None, presort='deprecated',
                           random_state=0, subsample=1.0, tol=0.0001,
                           validation_fraction=0.1, verbose=0,
                           warm_start=False)
```

```
[ ] pred_gbc = GBC.predict(xv_test)
```

```
[ ] GBC.score(xv_test, y_test)
```

```
0.8848484848484849
```

4. Random Forest Classifier

```
[ ] from sklearn.ensemble import RandomForestClassifier
```

```
[ ] RFC = RandomForestClassifier(random_state=0)
RFC.fit(xv_train, y_train)
```

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                        criterion='gini', max_depth=None, max_features='auto',
                        max_leaf_nodes=None, max_samples=None,
                        min_impurity_decrease=0.0, min_impurity_split=None,
                        min_samples_leaf=1, min_samples_split=2,
                        min_weight_fraction_leaf=0.0, n_estimators=100,
                        n_jobs=None, oob_score=False, random_state=0, verbose=0,
                        warm_start=False)
```

```
[ ] RandomForestClassifier(random_state=0)
```

```
RandomForestClassifier(bootstrap=True, ccp_alpha=0.0, class_weight=None,
                        criterion='gini', max_depth=None, max_features='auto',
                        max_leaf_nodes=None, max_samples=None,
                        min_impurity_decrease=0.0, min_impurity_split=None,
                        min_samples_leaf=1, min_samples_split=2,
                        min_weight_fraction_leaf=0.0, n_estimators=100,
                        n_jobs=None, oob_score=False, random_state=0, verbose=0,
                        warm_start=False)
```

```
[ ] pred_rfc = RFC.predict(xv_test)
```

```
[ ] RFC.score(xv_test, y_test)
```

```
0.9131313131313131
```

```
[ ] print(classification_report(y_test, pred_rfc))
```

	precision	recall	f1-score	support
0	0.91	0.91	0.91	241
1	0.92	0.91	0.92	254
accuracy			0.91	495
macro avg	0.91	0.91	0.91	495
weighted avg	0.91	0.91	0.91	495

Model Testing With Manual Entry

News

```
[ ] def output_lable(n):
    if n == 0:
        return "Fake News"
    elif n == 1:
        return "Not A Fake News"

def manual_testing(news):
    testing_news = {"article":[news]}
    new_def_test = pd.DataFrame(testing_news)
    new_def_test["article"] = new_def_test["article"].apply(wordopt)
    new_x_test = new_def_test["article"]
    new_xv_test = vectorization.transform(new_x_test)
    pred_LR = LR.predict(new_xv_test)
    pred_DT = DT.predict(new_xv_test)
    pred_GBC = GBC.predict(new_xv_test)
    pred_RFC = RFC.predict(new_xv_test)

    return print("\n\nLR Prediction: {} \nDT Prediction: {} \nGBC Prediction: {} \nRFC Prediction: {}".format(output_lable(pred_LR[0]),
                                                                 output_lable(pred_DT[0]),
                                                                 output_lable(pred_GBC[0]),
                                                                 output_lable(pred_RFC[0])))
```

```
[ ] news = str(input())
    manual_testing(news)
```

প্রেম করলে বাড়বে গুজন!

LR Prediction: Fake News
DT Prediction: Fake News
GBC Prediction: Fake News
RFC Prediction: Fake News

PLAGIARISM REPORT

AdilDFN

ORIGINALITY REPORT

9%	4%	3%	5%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Submitted to Dhirubhai Ambani Institute of Information and Communication Student Paper	2%
2	dspace.daffodilvarsity.edu.bd:8080 Internet Source	1%
3	Saif Mahmud Khan Dourjoy, Abu Mohammed Golam Rabbani Rafi, Zerine Nasrin Tumpa, Mohd. Saifuzzaman. "Chapter 49 A Comparative Study on Prediction of Dengue Fever Using Machine Learning Algorithm", Springer Science and Business Media LLC, 2021 Publication	1%
4	programmerbackpack.com Internet Source	1%
5	Submitted to Da Nang University of Economics Student Paper	<1%
6	Submitted to Amrita Vishwa Vidyapeetham Student Paper	<1%